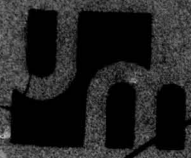
AD-A045 594 MISSOURI UNIV-COLUMBIA DEPT OF STATISTICS F/G 12/1
LEAST SQUARES VIEWED AS A GENERAL OPTIMIZATION PROBLEM.(U)
JUN 77 R P KELLEY N00014-75-C-0443
UNCLASSIFIED TR-68 NL

| OF |
AD
A045594

END
DATE FILMED
11-77
DDC



University of Missouri-Columbia

Least Squares Viewed as a General Optimization Problem

by

R. P. Kelley
University of Missouri-Columbia

Technical Report No. 68
Department of Statistics
June 1977

Mathematical Sciences

LEAST SQUARES VIEWED AS A GENERAL OPTIMIZATION PROBLEM.[1]

by

R. P. Kelley

University of Missouri-Columbia

ABSTRACT

Least squares problems arise when one attempts to fit a model $y = \eta(x,\beta)$ to points $(y_1, x_1),\ldots,(y_n, x_n)$. Solutions to such problems are obtained by minimizing the sum of squared deviations, over an admissible region. This paper discusses the basic theory of optimization for a general objective function and applies this material to both linear and nonlinear least squares problems. In linear least squares, normal equations for both the full rank and less than full rank cases are considered and the Kuhn-Tucker conditions are used to obtain the normal equations under linear inequality constraints. In nonlinear least squares different iterative methods which may be used to obtain a solution are discussed. The methods considered are steepest descent, Newton-Raphson, Gauss-Newton, Hartley's modified Gauss-Newton, and that of Marquardt. Results are obtained which relate Marquardt's method to equality constrained linear least squares.

[1]Partially supported by the Office of Naval Research under Contract N00014-75-C-0443. Reproduction in whole or part is permitted for any purpose of the United States Government.

1. OPTIMIZATION

1.1 Iteration Procedures

We are primarily concerned with optimizing, i.e. maximizing or minimizing, an objective function $f(t) = f(t_1,\ldots,t_p)$ over an admissible region $K \subset E^p$.

A general iteration procedure is as follows:

(i) choose a seed point $t \in K$,

(ii) on the basis of local behavior of $f(\cdot)$ at t select a new iterate $At \in K$,

(iii) replace t by At and return to step (ii).

A is a mapping of K into itself. The properties of such an iteration scheme are the properties of A.

The following are examples of some co[illegible]teration procedures on E^p.

Example 1. Steepest Descent

A descent direction, s, at the point t, is a unit vector such that

$$\frac{d}{d\lambda} f(t + \lambda s)\Big|_{\lambda=0} = s^T \nabla f(t) < 0;$$

hence, any descent direction must be within $\pi/2$ of the negative gradient, $-\nabla f(t)$.

A general descent procedure is:

(i) choose a seed point $t \in E^p$,

(ii) at t, select a descent direction s and a step length λ,

(iii) return to ii), replacing t by

$$At = t + \lambda s.$$

The steepest descent technique chooses as its descent direction, $-\nabla f(t)$. This is a locally optimal choice in that

$$\frac{d\, f\, (t + \lambda s)}{d\, \lambda}\Big|_{\lambda=o} = s^T \nabla f(t)$$

is minimized for s proportional to $-\nabla f(t)$

The corresponding technique for maximization is the method of steepest ascent. See Nobel (1969, p.403) for a discussion of convergence.

Example 2. Newton's Method

Newton's method is designed to solve $h(t) = 0$; where h is a continuously differentiable, real valued function of a real variable. The mapping which defines Newton's method, say Nt*, is obtained by solving $\frac{dh}{dt}(t^*)(t - t^*) + h(t^*) = 0$ for t. That is, we are to find the intercept of the tangent to the curve determined by h, at the point $(t^*, h(t^*))$; hence,

$$Nt = t - h(t) / \frac{dh(t)}{dt}$$

if $\frac{dh(t)}{dt} \neq 0$.

While Newton's method is designed specifically to solve nonlinear equations, it can be used to obtain a critical point of an objective function f and thus a potential solution to the optimization problem. Specifically, to solve $f'(t) = 0$ by Newton's method, we have

$$Nt = t - f'(t)/f''(t).$$

Example 3. Newton-Raphson method

The Newton-Raphson method is a p-dimensional analog of Newton's method. Let $t \in E^p$ and $h(t) = (h_1(t),\ldots,h_p(t))^T$; assume that h_k has continuous first order partials for $k = 1,\ldots,p$. For each k consider the hyperplane tangent to the $p + 1$ dimensional surface determined by h_k at the point $(t^*, h_k(t^*))$. The equations of these planes are

$$z = h_k(t^*) + (t - t^*)^T \nabla h_k(t^*): k = 1,\ldots,p. \tag{1.1}$$

We obtain Nt^* by finding the point of intersection of the tangent planes (1.1) with the plane $z = 0$. This yields

$$(Nt - t)^T \nabla h_k(t) = -h_k(t); \; k = 1,\ldots,p.$$

When the Newton-Raphson method is employed to find critical points of an objective function f we get the matrix equation

$$(Nt - t)^T \nabla^2 f(t) = -(\nabla f(t))^T,$$

where $\nabla^2 f(t)$, the Hessian, is

$$\nabla^2 f(t) = \left(\frac{\partial^2 f}{\partial t_k \partial t_e}\right).$$

The convergence of a general iteration procedure is of considerable importance. A condition on A sufficient to guarantee convergence is given by Kolmogorov and Fomin (1957, p. 43). Let R be a metric space with metric ρ. A mapping, A, of R into itself is a contraction if there exists α $(0 < \alpha < 1)$ such that

$\rho(At, At^*) \leq \alpha\rho(t, t^*)$ for all $t, t^* \in R$.

Theorem 1.1 (Principle of Contraction Mappings)

Every contraction mapping defined in a complete metric space R has one and only one fixed point. (i.e. the equation $At = t$ has one and only one solution).

Furthermore, Kolmogorov and Fomin's proof of this Theorem implies that, given a seed point t, the sequence $t, At, A^2t, \ldots$ converges to the fixed point.

Theorem 1.2

Let f be defined on [a, b]. If $f'(a) < 0 < f'(b)$ and $0 < k_1 \leq f''(t) \leq k_2$ on [a, b] then the equation $f'(t) = 0$ has a unique solution in (a, b) and the mapping $At = t - \lambda f'(t)$ is a contraction for $0 < \lambda < k_2^{-1}$.

Proof

f' is continuous and strictly increasing with $f'(a) < 0 < f'(b)$. Therefore $f'(t) = 0$ has a unique solution in (a, b).

If $t_0, t_1 \in [a, b]$ then

$$At_0 - At_1 = [1 - \lambda f''(\xi)](t_0 - t_1); \quad a \leq \xi \leq b.$$

In particular, for $t \in [a, b]$ and $0 < \lambda < k_2^{-1}$,

$$At - a = [1 - \lambda f''(\xi)](t - a) + Aa - a > 0.$$

Similarly $b > At$. Finally,

$$|At_0 - At_1| = |1 - \lambda f''(\xi)||t_0 - t_1| < (1 - \lambda k_1)|t_0 - t_1|.$$

Therefore A is a contraction on [a, b].

As a corallary we see that, under the conditions of the Theorem, the method of steepest descent (with fixed step length) converges to a unique minimum of f.

1.2 Optimization with Constraints

The mathematical programming problem is as follows: Given the numerical functions $f, g_1, \ldots, g_m$ defined on E^p, find a point t of E^p satisfying

$$g_j(t) \geq 0, \quad j = 1, \ldots, m$$

and such that f(t) is as large as possible. A solution of the problem will be denoted by $\hat{t}$. Minimization problems may be handled by taking $-f(\cdot)$ as the objective function.

The inequalities $g_j(t) \geq 0$ are called the constraints of the program; points which satisfy the constraints are feasible points and the set of feasible points is the feasible region, denoted by K. Throughout this chapter it will be assumed that f and $g_1, \ldots, g_m$ are differentiable.

Two well-known special cases are i) when f and $g_1, \ldots, g_m$ are linear, we have the linear programming problem and ii) when f is a quadratic form and $g_1, \ldots, g_m$ are linear, we have the quadratic programming problem.

Kuhn and Tucker (1951) developed a set of conditions, the K-T conditions which, under mild regularity conditions are necessary for a solution to the programming problem. Their conditions are that there exist $u_1, u_2, \ldots, u_m$ such that

$$g_j(t) \geq 0; \quad j = 1,\ldots, m$$

$$u_j \geq 0; \quad j = 1,\ldots, m$$

$$u_j \cdot g_j(t) = 0; \quad j = 1,\ldots, m$$

and

$$\nabla f(t) + \sum_{j=1}^{m} u_j \nabla g_j(t) = 0.$$

There are several sets of such regularity conditions which can be employed. Here we do not try for the most general results but we give conditions, involving generalized concavity, which are easily understood and yet quite general.

Differentiable functions with the property that

$$f(y) > f(x) \quad \text{implies} \quad \nabla f(x) \cdot (y - x) > 0$$

(increasing function implies positive slope) are called pseudo-concave by Mangasarian (1965).

In his unpublished 1953 notes, Convex Cones, Sets and Functions, W. Fenchel treats the concept of quasi-concavity. A real valued function $f(x)$ having convex domain is called quasi-concave (q-concave) if $f(\lambda x + \bar{\lambda} y) \geq \min(f(x), f(y))$ whenever $0 < \lambda < 1$ and $\bar{\lambda} = 1 - \lambda$. For differentiable functions, pseudo-concave implies q-concave.

Several alternative characterizations of q-concave functions are available.

Theorem 1.3

The following conditions are equivalent

(i) $f(x)$ is q-concave

(ii) $I_\tau = \{x : f(x) > \tau\}$ is convex for all τ,

(iii) the level sets $L_\tau = \{x : f(x) \geq \tau\}$ are convex for all τ.

The alternative characterizations of Theorem 1.3 are important; we see for example, that convexity of the "constraint set" $\{x: g_1(x) \geq \tau_1, g_2(x) \geq \tau_2, \ldots g_m(x) \geq \tau_m\}$ is assured for all $\tau_1, \ldots, \tau_m$ when and only when $g_1, g_2, \ldots, g_m$ are q-concave functions.

We may now state the following results, see Mangasarian (1969).

Theorem 1.4

If f $, \ldots, g_m$ are differentiable, the g_j's are pseudo-concave [illegible] there exists some feasible a such that $g_j(a) > 0$ for all g_j which are not affine, and if $\hat{t}$ is a solution of the programming problem then there exists $u = (u_1, \ldots, u_m)^T$ such that $\hat{t}$ and u satisfy the K-T conditions.

Generalized concavity can also provide a framework within which the K-T conditions are sufficient.

Theorem 1.5

Let f be a differentiable and pseudo-concave function and $g_1, \ldots, g_m$ be differentiable and quasi-concave. If t and u solve the Kuhn-Tucker conditions then t solves the mathematical programming problem; that is $t = \hat{t}$.

As an example consider optimizing a quadratic objective function subject to linear inequality constraints. This is often called quadratic programming. In general, we would have the following problem:

Minimize $\frac{1}{2} x^T F x - x^T d$

Subject to $G^T x \geq b$

where F and G are matrices, with F symmetric, while b and d are vectors.

The K-T conditions,

$$G^T x \geq b, \; u \geq 0$$

$$u^T (G^T x - b) = 0$$

$$Gu = Fx - d$$

are necessary for a solution to the problem (see Theorem 1.4). If F is positive semidefinite then, the objective function is concave and Theorem 1.5 tells us that the K-T conditions are also sufficient.

We have

$$\frac{1}{2} x^T F x - x^T d = \frac{1}{2}(x - x)^T F (x - c) + \text{constant}$$

for some c, if and only if $x^T d = x^T F c$ for all x. This last is equivalent to the equation $Fc = d$ which has a solution, c, when and only when d is in the column space of F. Thus the quadratic programming problem can be written in the special form: minimize $\frac{1}{2}(x - c)^T F (x - c)$ suject to $G^T x \geq b$ if and only if d is in the column space of F. In particular, the positive definite quadratic programming problem can always be written in this form.

2. LEAST SQUARES

2.1 Linear Least Squares

Least Squares problems arise when one attempts to fit a model $y = \eta(x,\beta)$ to points $(y_1, x_1),\ldots, (y_n, x_n)$. Here η is a function of known form, β is a parameter to be estimated, and x is a vector of independent variables. We obtain the fitted model by minimizing

$$\phi(\beta) = \sum_{i=1}^{n} [y_i - \eta(x_i,\beta)]^2 \tag{2.1}$$

with respect to β, a solution being denoted by $\hat{\beta}$.

The <u>deviations</u> $d_i(\beta) = y_i - \eta(x_i, \beta)$; $i = 1,\ldots, m$, or $d = Y - \eta(\beta)$, where $\eta(\beta) = \big(\eta(x_1,\beta),\ldots, \eta(x_n,\beta)\big)^T$, measure the goodness of fit of the model $y = \eta(x,\beta)$ at the parameter value β. The deviations $d_i(\hat{\beta})$; $i = 1,\ldots, n$ are called the <u>residuals</u>.

If $\eta(x,\beta) = x\beta$ then we have the special case of linear least squares, and (2.1) becomes

$$\phi(\beta) = \sum_{i=1}^{n} [y_i - x_i\beta]^2 \ . \tag{2.2}$$

Let X be the matrix whose ith row is x_i, $Y = (y_1 \ldots y_n)^T$, and V be the column space of X, then by minimizing (2.2) we are finding a $\hat{\beta}$ such that $X\hat{\beta}$ is the vector in V that is the closest to Y. Thus, $\hat{\beta}$ solves the linear least squares problem if and only if $X\hat{\beta}$ is the projection of Y on to V. From the projection theorem, $X\hat{\beta}$ is that vector such that we may write $Y = X\hat{\beta} + (Y - X\hat{\beta})$ with $X\hat{\beta} \in V$ and $y - X\hat{\beta}$ orthogonal to V; hence, $X^T(Y-X\hat{\beta}) = 0$ or

$$X^TX\hat{\beta} = X^TY. \tag{2.3}$$

The equations (2.3) are called the normal equations.

We have seen that if $\hat{\beta}$ solves the linear least squares problem then it solves the normal equations. Suppose that $\hat{\beta}$ satisfies the normal equations, then $X^TX\hat{\beta} = X^TY$; thus $X^T(Y - X\hat{\beta}) = 0$. Therefore $X\hat{\beta}$ is the projection of Y on V; hence, $\hat{\beta}$ solves the linear least squares problem. Thus $\hat{\beta}$ solves the linear least squares problem iff $\hat{\beta}$ solves the normal equations.

If X is of full rank then X^TX is nonsingular; hence, the normal equations have a unique solution

$$\hat{\beta} = (X^TX)^{-1} X^TY.$$

However, if X is not of full rank then there will exist infinitely many solutions; a general solution to (2.3) will be given by

$$\hat{\beta} = (X^TX)^-X^TY + \left((X^TX)^-(X^TX) - I\right)Z,$$

where Z is arbitrary and $(X^TX)^-$ denotes the generalized inverse of X^TX, see Rao (1965, p. 24).

We now briefly consider linear least squares with constraints. To this end we first reformulate the unconstrained problem:

Minimize $(Y - \eta)^T(Y - \eta)$ subject to $\eta = X\beta$ or $\eta \in V$. Alternatively we may take the feasible region to be the space orthogonal to the space orthogonal to V. Let $P_{r+1}, \ldots, P_n$ be a basis for the space orthogonal to V and let $G^T = (P_{r+1}, \ldots, P_n, -P_{r+1}, \ldots, -P_n)$. The constraints can be written $G^T\eta \geq 0$.

If in addition we have the constraints

$$W^T\eta \geq b$$

then our problem becomes

$$\text{minimize } (Y - \eta)^T (Y - \eta)$$

subject to $W^T\eta \geq b$ and $G^T\eta \geq 0$.

The K - T conditions are:

$$\eta = X\beta$$

$$Q \geq 0 , \quad W^T\eta \geq b$$

$$Q^T(W^T\eta - b) = 0$$

and

$$\sum_{i=r+1}^{n} P_i u_i + WQ = \eta - Y$$

where u_i is unrestricted in sign.

Replacing η by $X\beta$ and multiplying by X^T we obtain

$$\sum_{i=r+1}^{n} X^T P_i \mu_i + X^T WQ = X^T X \beta - X^T Y.$$

or

$$X^T X \beta = X^T (Y + WQ).$$

These are the generalized normal equations to be solved for β and Q.

The following results are of interest for nonlinear estimation, see Marquardt (1963). However, they are actually theorems in constrained linear least squares, so we present them here and return to them later. Consider the programming problem:

$$\text{minimize } ||Y - X\beta||, \text{ with respect to } \beta,$$

(2.4)

$$\text{subject to } ||\beta|| \leq r.$$

The Kuhn-Tucker conditions are necessary and sufficient for this program; this gives the following result.

Theorem 2.1

β solves the program (2.4) if and only if either

$$\text{a)} \quad X^T X\beta = X^T Y \text{ and } ||\beta|| \leq r$$

or

b) there exists $u > 0$ such that

$$(X^T X + uI)\beta = X^T Y, \quad ||\beta|| = r .$$

Theorem 2.2

Let u and β_u satisfy $u > 0$ and $(X^T X + uI)\beta_u = X^T Y$. $||\beta_u||$ is a strictly decreasing function of u approaching 0 as u tends to ∞.

Proof

There exists an orthogonal matrix S such that $S^T X^T X S = D = \text{diag}(d_1, \ldots, d_p)$. Since $u > 0$, $\beta_u = S^T(D + uI)^{-1} SX^T Y$. Writing $SX^T Y = V = (v_1, \ldots, v_p)^T$ we get

$$||\beta_u||^2 = V^T(D + uI)^{-2}V = \sum_{j=1}^{p} \left(\frac{v_j}{d_j + u}\right)^2,$$

from which the truth of the theorem is evident.

Theorem 2.3

Let α be the angle between β_u and X^TY then α is strictly decreasing as a function of u and α tends to 0 as u tends to ∞.

Proof

Since $0 \leq \alpha \leq \Pi$, we may instead prove that $\cos \alpha$ is a strictly increasing function of u and that $\cos \alpha$ tends to 1 as u approached ∞.

$$\cos \alpha = \frac{Y^TXS^T(D + uI)^{-1}SX^TY}{(Y^TXS^T(D + uI)^{-2}SX^TY)^{\frac{1}{2}}(Y^TXX^TY)^{\frac{1}{2}}}$$

$$= \frac{\sum_{j=1}^{p} \frac{v_j^2}{d_j+u}}{\left(\sum_{j=1}^{p} \frac{v_j^2}{(d_j+u)^2}\right)^{\frac{1}{2}} ||X^TY||} .$$

As $u \to \infty$ we see that

$$\cos \alpha \to \frac{\left(\sum_{j=1}^{p} v_j^2\right)^{\frac{1}{2}}}{||X^TY||} = 1.$$

Also,

$$\frac{d \cos \alpha}{d u} = \frac{\left(\sum_j \frac{v_j^2}{d_j+u}\right)\left(\sum \frac{v_j^2}{(d_j+u)^3}\right) - \left(\sum \frac{v_j^2}{(d_j+u)^2}\right)^2}{\left(\sum_j \frac{v_j^2}{(d_j+u)^2}\right)^{3/2} ||X^TY||} ;$$

using Swartz's inequaltiy, $d \cos \alpha / du > 0$.

Related to the above two results is the expansion

$$\beta_u = u^{-1}X^TY - u^{-2}(X^TX)X^TY + u^{-3}(X^TX)^2X^TY - \cdots \tag{2.5}$$

valid for u greater than the maximum characteristic root of X^TX. This is obtained from the geometric expansion for matrices

$$(M + I)^{-1} = I - M + M^2 - \cdots$$

, see for example Friedman (1956).

2.2 Nonlinear Least Squares

Nonlinear least squares problems arise when one attempts to fit a model $y = \eta(x,\beta)$ with η nonlinear in β.

We first make a general observation about residuals. If $\eta(x,\beta)$ is of the form $\eta(x,\beta) = \beta_1 + \psi(x; \beta_2, \ldots, \beta_p)$ then

$$\frac{\partial \phi\ (\hat{\beta})}{\partial\ \beta_1} = -2 \sum_{i=1}^{n} [y_i - \hat{\beta}_1 - \psi(x_i; \hat{\beta}_2, \ldots, \hat{\beta}_p)].$$

Equating this to 0 we get $\sum_{i=1}^{n} d_i(\hat{\beta}) = 0$; that is, the residuals sum to 0.

Explicit solutions will usually not be available in the nonlinear case and one must resort to iterative minimization techniques. We now present four iteration methods specifically adapted to the nonlinear least squares problem.

Steepest descent

The gradient of (2.1) is

$$\nabla\phi = -2X(\beta)^T[y - \eta(\beta)]$$

where

$$X(\beta) = \begin{pmatrix} \frac{\partial \eta(x_1, \beta)}{\partial \beta_1} & \cdots & \frac{\partial \eta(x_1, \beta)}{\partial \beta_p} \\ \vdots & & \vdots \\ \frac{\partial \eta(x_n, \beta)}{\partial \beta_1} & \cdots & \frac{\partial \eta(x_n, \beta)}{\partial \beta_p} \end{pmatrix} .$$

Hence the method of steepest descent specifies the mapping

$$S\beta = \beta + \lambda X(\beta)^T d(\beta)$$

but does not specify the step length, λ.

Gauss Newton method

The Gauss-Newton method is an iteration procedure which assumes local linearity of $\eta(x,\cdot)$ about β to obtain the new iterate $G\beta$. The equation of the tangent plane to the surface determined by $\eta(x,\cdot)$, at the point β^*, is

$$y = \eta(x,\beta^*) + \sum_{k=1}^{p} \frac{\partial \eta(x,\beta^*)}{\partial \beta_k}(\beta_k - \beta_k^*) .$$

Replacing η by its linear approximation reduces the problem to the previously considered case of linear least squares. That is we wish to minimize

$$\sum_{i=1}^{n} [y_i - \eta(x_i, \beta^*) - \sum_{k=1}^{p} \frac{\partial \eta(x_i,\beta^*)}{\partial \beta_k}(\beta_k - \beta_k^*)]^2$$

$$= ||y - \eta(\beta^*) - X(\beta^*)(\beta - \beta^*)||^2$$

with respect to β, or $||d(\beta^*) - X(\beta^*)\delta||^2$ with respect to

$\delta = \beta - \beta^*$. This is equivalent to projecting $d(\beta^*)$ onto the column space of $X(\beta^*)$. The solution is given by solving the normal equations

$$X(\beta^*)^T X(\beta^*)\delta = X(\beta^*)^T d(\beta^*).$$

Replacing β^* by β and δ by $G\beta - \beta$ we get

$$G\beta = \beta + [X(\beta)^T X(\beta)]^- X^T(\beta) d(\beta).$$

There exist examples with well behaved functions $\eta(x_i, \beta)$ for which the Gauss-Newton iteration will not converge no matter how good the starting value. However, Jennrich (1969) gives four conditions which are collectively sufficient so that such difficulties are not likely to arise when n, the sample size, is large and the starting value is close to the true parameter value, β.

One of the sufficient conditions just mentioned is that the parameter space is a compact subset of Euclidean space. Hence, in using the results obtained one would want to restrict the investigation to some closed and bounded subset of the parameter space.

Hartley's Modified Gauss-Newton Method

In considering the Gauss-Newton method we notice that, given β, there is input of information from the objective function, ϕ, concerning the choice of the next iterate only through a quadradic approximation and it is possible for the value of the objective function to actually increase by iteration. This increase would not, in itself, invalidate the procedure but it could cause slow convergence. Hartley (1961) modifies the Gauss-Newton method to allow

greater input from the objective function in the iteration procedure. This input is obtained by making the following assumptions:

(a) the parameter space, Ω, is a convex subset of E^p;

(b) $\frac{\partial \eta(x_i,\beta)}{\partial \beta_k}$, $\frac{\partial^2 \eta(x_i,\beta)}{\partial \beta_\ell \partial \beta_k}$ exist and are continuous for all $i=1,\ldots, n$ and $\ell, k=1,\ldots, p$;

(c) there exists a bounded convex subset, S, of the parameter space such that for every $\beta \in S$ and every $\mu = (\mu_1,\ldots, \mu_p)' \neq 0$, $\sum_{i=1}^{n} \left(\sum_{r=1}^{p} \mu_k \frac{\partial \eta(x_i,\beta)}{\partial \beta_k} \right)^2 > 0$ (this is equivalent to requiring that $X(\beta)$ be of full rank in S);

(d) there exists β^0 in the interior of S such that $\phi(\beta^0) < \inf_{\beta \in S^c} \phi(\beta)$.

Hartley's algorithm is as follows:

(i) choose $\beta = \beta^0$ as starting vector;

(ii) obtain another estimate $G\beta$ by the usual Gauss-Newton method (the existance of $G\beta$ is guaranteed by assumption (c));

(iii) consider $\phi(\lambda\beta + (1-\lambda)G\beta)$, $\lambda \in [0, 1]$, and let $\lambda^* \in [0, 1]$ be such that $\min_{\lambda \in [0,1]} \phi(\lambda\beta + (1-\lambda)G\beta) = \phi(\lambda^*\beta + (1-\lambda^*)G\beta)$ (from (b) $\phi(\beta)$ is continuous and hence $\phi(\lambda\beta + (1-\lambda)G\beta)$

obtains a minimum on [0, 1]);

(iv) replace β by $H\beta = \lambda^*\beta + (1-\lambda^*)G\beta$ and return to (ii).

Hartley argues that given a sequence, $\{H^j\beta^0\}$, constructed by this algorithm, there exists a subsequence, $\{H^{j_k}\beta^0\}$, converging to a point $\hat{\beta}$ which is a critical point of $\phi(\beta)$. If, in addition to the above assumptions, the Hessian of $\phi(\beta)$ is positive definite on S, then $\hat{\beta}$ is the unique minimum of $\phi(\beta)$.

The Marquardt method

Marquardt's algorithm for the solution of nonlinear least squares problems is a compromise between the Gauss-Newton and steepest descent methods, the objective of this compromise being the avoidance of problems associated with the two methods.

Let us first review the major difficulties attributed to the use of the Gauss-Newton and steepest descent methods. First, steepest descent does not specify the step length. Second, if the level sets of ϕ tend to be elongated then the method of steepest descent will converge rapidly for the first few iterates and then oscillate about the axis of elongation taking smaller and smaller steps as the iterates approach the minimum. This is because the correction for β obtained in steepest descent is perpendicular to the level set at β; hence, for points close to the minimum, the correction vector may be almost perpendicular to the direction of the minimum if the level sets are greatly elongated. The main problem encountered with the Gauss-Newton method is lack of convergence of the iteration if the starting points are a long way from the minimum.

One possible solution to these problems is to use the steepest descent method for the first few iterations and then switch to the Gauss-Newton method when the progress becomes slow. Marquardt's method is one way in which this can be done.

As in the Gauss-Newton method, assume local linearity of $\eta(x;\cdot)$ at the point β. Theorem 2.1 then states that the issue concerning the choice of step length can be restated in terms of a La Grange multiplier u. More specifically, least squares subject to a constraint on the maximum step length leads to the equation

$$[X^T(\beta)X(\beta) + uI]\delta_u = X^T(\beta)d(\beta)$$

where δ_u is the correction to β and $d(\beta) = Y - \eta(\beta)$. From (2.5) we get, for large u,

$$\delta_u \simeq u^{-1} X^T(\beta)[Y - \eta(\beta)]$$

But $x^T(\beta)[Y - \eta(\beta)]$ is the direction of steepest descent and ϕ is continuous so that for u sufficiently large

$$\phi(\delta_u + \beta) < \phi(\beta) \tag{2.6}$$

Marquardt (1963) recommends that the next iterate say $M\beta$ be given by $M\beta = \beta + \delta_u$ where u is chosen just large enough to satisfy (2.6).

Thus, in outline, δ_o is the correction given by the Gauss-Newton method while for large u, δ_u is in the direction of steepest descent. Thus $\beta + \delta_u$ determines a continuous curve on which Marquardt's method interpolates between the Gauss-Newton and steepest descent methods.

This manuscript is the joint work of the author and W. A. Thompson, Jr.

BIBLIOGRAPHY

Friedman, B. (1956). Principles and Techniques of Applied Mathematics, Wiley.

Hartley, H. O. (1961). The modified Gauss-Newton method, Technometrics, Vol. 3, #2, May pp. 269-280.

Jennrich, R. I. (1969). Asymptotic properties of nonlinear least squares estimators, Annals of Math Stat., Vol. 40, #2 pp. 633-643.

Kolmogorov, A. N. and Fomin, S. V. (1957). Elements of the Theory of Functions and Functional Analysis, Graylock.

Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear Programming, Proceedings of the Second Berkeley Symposiumn on Mathematical Statistics and Probability, University of California Press, Berkeley, pp. 481-492.

Mangasarian, O. L. (1965). Pseudo-convex functions, J. Soc. Indust. Appl. Math, Vol # 3, pp. 281-290.

Mangasarian, O. L. (1969). Nonlinear Programming, McGraw-Hill.

Marquardt, D. W. (1963). An algorithm for least squares estimation of nonlinear parameters, J. Soc. Indust. Appl. Math., Vol. 11, #2, June, pp. 431-441.

Nobel, B. (1969). Applied Linear Algebra, Prentice-Hall.

Rao, C. R. (1965). Linear Statistical Inference, Wiley.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 77-1
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): Least Squares Viewed as a General Optimization Problem.
5. TYPE OF REPORT & PERIOD COVERED: Technical rept.
6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): R. P. Kelley
8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0443 (NR042-282)

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, University of Missouri-Columbia
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS: TR-68, 77-1
12. REPORT DATE: 16 June 1977
13. NUMBER OF PAGES: 24p.

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Constraints, Iteration, least squares, optimization

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Least squares problems arise when one attempts to fit a model $y = \eta(x,\beta)$ to points $(y_1,x_1),\ldots,(y_n,x_n)$. Solutions to such problems are obtained by optimizing the sum of squared deviations over an admissible region. This paper discusses the basic theory of optimization for a general objective function and applies this material to both the linear

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. and nonlinear least squares problems.

In linear least squares normal equations for both the full rank and less than full rank cases are considered and the Kuhn-Tucker conditions are used to obtain the normal equations under linear inequality constraints. In non-linear least squares, different iterative procedures, which may be used to obtain a solution, are discussed. The methods considered are steepest descent, Newton-Raphson, Gauss-Newton, Hartley's modified Gauss-Newton, and that of Marquardt. Results are obtained which relate Marquardt's method to equality constrained least squares.